

Open in app ↗

Sign up

Sign in

Medium

🔍 Search



Amazon's \$10 Billion Bet That Nobody Can Copy Its AI Chips

9 min read · Just now



Berend Watchus

Follow



Listen



Share

Author: Berend Watchus December 4, 2025

Trade Secrets vs. Patents in the Ultimate Speed Game



logo copyright: Amazon

About Amazon

New AWS AI Factories transform customers' existing infrastructure into high-performance AI environments

AWS AI Factories deliver dedicated infrastructure combining the latest NVIDIA accelerated computing platform, Trainium chips, AWS AI...

2 dagen geleden



TechCrunch

Amazon releases an impressive new AI chip and teases an Nvidia-friendly roadmap

Amazon Web Services, which has been building its own AI training chips for years now, just introduced a new version known as Trainium3 that...

2 dagen geleden



Barron's

Amazon Is Taking On Nvidia and Google. Why It's a Big Deal for the Stock.

Amazon said "UltraServers" powered by its in-house Trainium 3 chips are now available, looking to rival Nvidia and Google.

2 dagen geleden



(Amazon Trainium chips, world news)

Amazon's \$10 Billion Bet That Nobody Can Copy Its AI Chips

How Trainium3 Is Protected by Secrecy, Not Patents — And Why Amazon Thinks That's Enough

LAS VEGAS, Dec 4, 2025 — Two days ago at re:Invent, AWS unveiled Trainium3: the world's first commercial 3 nm AI training chip. The numbers are brutal: 4.4× the performance of Trainium2, 40% better efficiency, 144 chips per UltraServer, already training the next Claude models for Anthropic, and promising customers 50% lower training and inference costs.

What almost nobody in the keynote mentioned is this: **almost none of the technology that actually delivers those gains is meaningfully protected by patents.**

Search the USPTO, Google Patents, or Justia today for anything that describes Trainium3's new systolic-array density tricks, the stochastic-rounding units, the NeuronSwitch-v1 all-to-all fabric, or exactly how they doubled MXFP8 throughput on TSMC's N3P node. You will find nothing specific after 2023 and nothing that would stop a determined competitor from copying the design tomorrow.

Amazon has invested an estimated **\$3–8 billion** (silicon design, mask sets, early production runs, and the Alchip/TSMC manufacturing relationship) on a product whose crown jewels are defended almost entirely by corporate secrecy, not by the patent system.

And we are not talking about black-triangle stealth aircraft locked away in Area 51 hangars. We are talking about a **physical, commercial product** that AWS will ship by the hundreds of thousands into public data centers, sell through normal EC2 channels, and even offer as on-premise "AWS Outposts" racks to any enterprise or government that wants one.



× 🔊 🔍

AI-modus Alle Afbeeldingen Video's **Nieuws** Boeken Maps Meer ▾ Tools ▾


About Amazon

Frontier agents, Trainium chips, and Amazon Nova: key announcements from AWS re:Invent 2025

AWS introduces Graviton5 processors, its most powerful CPU, alongside Amazon Bedrock AgentCore, Trainium3 UltraServers, and AI Factories.

1 dag geleden




mint

Samsung officially unveils Exynos 2600 chip ahead of Galaxy S26 series release

Samsung officially unveiled the Exynos 2600, the world's first 2nm mobile chip, promising top-tier performance and efficiency for the...

8 uur geleden




Reuters

Amazon to use Nvidia tech in AI chips, roll out new servers

Amazon.com's AWS cloud computing unit on Tuesday said it will adopt key Nvidia technology in future generations of its artificial...

2 dagen geleden



world news

The Exposure Problem

The moment the first Trn3 UltraServer ships in Q1 2026, **anyone on Earth with a credit card or a purchase order** can legally and logistically acquire it:

- **NVIDIA's** silicon reverse-engineering lab in Santa Clara
- **Google's** TPU team in Mountain View
- **Meta's** MTIA group in Menlo Park
- **Microsoft's** Maia team in Redmond
- **Biren, Cambricon, Huawei HiSilicon**, or any Chinese state-backed institute through gray-market resellers in Singapore or Dubai
- **University cleanrooms** in Tehran, Moscow, Pyongyang, or Tel Aviv
- **Independent teardown houses** like TechInsights, ChipRebel, or System Plus Consulting

They can bolt it into a rack, delid it the same afternoon, and start the FIB-SEM cameras rolling. No export license currently required (Trainium3 is not yet on the Commerce Control List, though this will likely change given geopolitical tensions around AI hardware), no security clearance, no special NDA beyond the standard EC2 terms of service.

That is how completely exposed the hardware actually is.

This Is Not How Most Industries Work

When Boeing invents a new composite wing spar or Toyota develops a solid-state battery, they file hundreds of narrow, defensive patents — because they know Airbus or Tesla can buy the plane or the car, tear it apart, and copy it in a few years.

Yet Amazon is betting tens of billions of future AWS revenue that none of the five best-funded silicon teams on the planet will be able to do the same thing with Trainium3.

The Five Players Who Can Buy Trainium3 Tomorrow

1. **NVIDIA** — World's leading GPU manufacturer with extensive reverse-engineering capabilities
2. **Google (Tensor/TPU)** — Creator of the original TPU architecture, deep ML hardware expertise
3. **Meta (MTIA)** — Building custom AI training chips, significant semiconductor investment
4. **Microsoft (Maia)** — Developing proprietary AI accelerators, tight OpenAI integration
5. **China (Biren, Cambricon, Huawei, state labs)** — Nation-state backed with billions in funding

All five have reverse-engineering labs that are world-class. All five have 3 nm PDKs from TSMC or Samsung. All five have billions to burn.

So Why Is Amazon So Confident This Won't Happen in 12–36 Months?

1. 3 nm Is Still a Duopoly

TSMC and Samsung will simply refuse to tape-out an obvious Trainium3 clone. Losing Amazon as a top-5 customer (and risking U.S. sanctions) is not worth it. The leading-edge semiconductor manufacturing market is so concentrated that foundries can effectively police their customer base. A blatant copy would immediately raise red flags during the design submission process.

2. Full Reverse-Engineering Takes 4–8 Years and \$1–3 Billion

Even the best labs quote 4+ years and >\$100 million just for imaging and delayering a 3 nm chip. Modern chips have 15–20 metal layers, billions of transistors, and feature sizes approaching atomic limits. Turning that into clean, timing-closed RTL on a different PDK adds another 2–4 years and another billion. By then Trainium7 will be shipping, and the competitive window will have closed.

The economics simply don't work: spend \$3 billion over 6 years to copy a chip that

will be obsolete before you finish.

3. The Real Magic Lives in Closed-Source Software

Even a perfect silicon clone is useless without the Neuron compiler, runtime, collective libraries, and decades of PyTorch integration that deliver the last 40% of real-world performance. Amazon has invested heavily in the software stack that sits between TensorFlow/PyTorch and the hardware. This includes:

- Custom kernel libraries optimized for Trainium's architecture
- Distributed training frameworks tuned for NeuronLink interconnect
- Memory management systems that exploit HBM3e bandwidth
- Model compilation pipelines that map neural networks to systolic arrays

Hardware without software is just expensive sand. The Neuron SDK represents thousands of engineer-years of optimization work that can't be reverse-engineered from a silicon die.

4. The Die Is Actively Obfuscated

Dummy poly, scrambled standard cells, heavy OPC tricks — modern AI chips are designed so automated cell recognition fails on >60% of the layout. Human reverse-engineers have to hand-trace critical paths for years. Amazon (through Annapurna Labs) has incorporated anti-reverse-engineering techniques directly into the physical design:

- Intentionally complex routing that obscures functional blocks
- Dummy structures that appear functional but do nothing
- Non-standard cell libraries that defeat automated recognition tools
- Strategic placement of decoy circuits

This isn't paranoia — it's standard practice for high-value custom silicon. The goal isn't to make reverse engineering impossible, just impractically expensive and slow.

5. Hyperscale Lock-In

By the time a competitor has a 70% accurate clone on 5 nm, the biggest models (Claude, Llama 4, Gemini 3) will have been retrained on Trainium with custom kernels. Switching costs become astronomical. Model training isn't just about raw compute — it's about the entire ecosystem:

- Training recipes optimized for specific hardware characteristics
- Checkpoint formats tuned to particular memory hierarchies
- Custom mixed-precision strategies that exploit Trainium's data types
- Distributed training patterns designed around NeuronLink topology

Once a foundation model provider has invested months and millions training on Trainium, switching to a competitor's hardware means starting over from scratch.

The Patent Reality: Strategic, Not Comprehensive

While the article's core thesis holds — Trainium3 is primarily protected by secrecy — Amazon does maintain a patent portfolio, though it's far more limited than most people assume:

Get Berend Watchus's stories in your inbox

Join Medium for free to get updates from this writer.

Enter your email

Subscribe

What patents likely exist:

- Core architectural concepts (systolic array variations, specific interconnect topologies)

- Novel circuit designs (custom floating-point units, stochastic rounding implementations)
- System-level innovations (UltraServer configurations, memory hierarchies)

What's NOT patented:

- Exact implementation details
- Specific performance optimizations
- Manufacturing process adaptations
- Software-hardware co-optimization techniques

The patents serve primarily as legal speed bumps for startups and smaller players, not as comprehensive protection against well-funded competitors. This is a calculated trade-off: patents require full disclosure of how something works, which can actually help competitors. Trade secrets never expire and never require disclosure.

Think of it like Formula 1 racing teams protecting their aerodynamic designs. While they file patents on some general innovations, the specific computational fluid dynamics models, manufacturing techniques, and setup parameters that actually deliver lap time advantages are kept as closely guarded trade secrets. Competitors can photograph the cars from every angle, but replicating the performance requires understanding thousands of subtle design choices that aren't visible or documented anywhere.

Amazon has chosen this racing team strategy for Trainium.

Historical Precedent: It's Worked So Far

This strategy isn't new for AWS:

- **Graviton1 (2018)** — 7 years later, no successful commercial clones despite being based on standard ARM cores
- **Inferentia1 (2019)** — 6 years later, unique architecture still uncopied

- **Trainium1 (2022)** — 3 years in, no competing implementation has emerged

The pattern is clear: AWS's custom silicon strategy, backed by Annapurna Labs' expertise, has successfully defended market position through speed and secrecy rather than patents.

Trainium3 is the biggest, most expensive test yet of whether this approach scales to cutting-edge 3 nm technology and trillion-parameter model training.

The Real Moat: Continuous Innovation Velocity

Amazon isn't trying to protect Trainium3 forever. They're trying to protect it long enough to ship Trainium4, then Trainium5, then Trainium6. The moat isn't the chip itself — it's the **organizational capability to design, manufacture, and deploy new chips faster than competitors can copy old ones.**

This is the flywheel:

1. Ship new chip generation every 18–24 months
2. Competitors take 4–6 years to reverse-engineer and copy
3. By the time clone ships, you're 2–3 generations ahead
4. Customers stay because switching costs are too high
5. Revenue funds next generation, accelerating the cycle

It's not about having an unbreakable lock. It's about running fast enough that the lock doesn't need to be unbreakable.

Why This Matters Beyond AWS

Amazon's approach represents a fundamental shift in how semiconductor IP can be protected in an era where:

- Leading-edge manufacturing is concentrated in two companies (TSMC, Samsung)
- Software differentiation matters as much as hardware

- Time-to-market beats perfection
- Ecosystem lock-in creates switching costs
- Nation-states are willing to invest billions in chip cloning programs

The traditional patent playbook — file broadly, defend aggressively, license strategically — may no longer be optimal for rapidly-evolving technology products. Speed, secrecy, and software integration may provide better protection than legal rights.

This has implications for:

- **Startups:** Don't assume patents will protect you from well-funded copiers
- **Investors:** Value organizational velocity and software moats, not just patent counts
- **Policy makers:** Export controls may matter more than patent enforcement
- **Competitors:** Copying the hardware is the easy part; replicating the ecosystem is impossible

The Geopolitical Wild Card

The article notes that Trainium3 is “not yet on the Commerce Control List” — but this is almost certain to change. Recent U.S. export restrictions have targeted:

- Advanced GPUs (NVIDIA H100/H200 restricted to China)
- Chip manufacturing equipment (Applied Materials, ASML restrictions)
- AI software (licensing requirements for certain applications)

Trainium3, with its 3 nm process and frontier AI training capabilities, will likely face export restrictions within 12–18 months. This adds another layer of protection beyond secrecy: legal prohibitions on sales to certain countries or entities.

However, export controls are notoriously leaky. Gray market channels, shell

companies, and academic exemptions mean that motivated state actors will eventually acquire Trainium3 hardware regardless of official restrictions.

The real protection remains speed: by the time China successfully reverse-engineers and manufactures Trainium3 on domestic fabs (likely 2029–2031), Amazon will be shipping Trainium7 or 8.

The Bottom Line

Amazon is not protecting Trainium3 like Boeing protects a jet engine or Apple protects an iPhone camera.

Amazon is protecting it like **Ferrari protects its engine management software and combustion chamber designs**: by keeping the critical implementation details as trade secrets while selling the finished product to anyone who can afford it. You can buy a Ferrari, tear it completely apart, measure every dimension, analyze every material — but replicating the performance requires understanding ten thousand engineering decisions that aren't documented anywhere and took decades of accumulated expertise to develop.

The patent portfolio is mostly a legal speed bump for startups and a minor annoyance for well-funded competitors. The real moat is pure, old-fashioned secrecy — plus the knowledge that the companies who could copy it fastest are too busy funding their own multi-billion-dollar roadmaps to play catch-up to AWS.

So far, that strategy has worked since Graviton1 in 2018 and Trainium1 in 2022.

Trainium3 is the biggest, most expensive test yet of whether secrecy alone can still defend a semiconductor franchise when your customers are training trillion-dollar models and your competitors have nation-state budgets.

Amazon is betting it can.

History says they might be right.

This analysis is based on public information about AWS Trainium3 announced at re:Invent 2025, patent database searches conducted December 2025, and industry knowledge of semiconductor reverse-engineering timelines and costs.

Trade Secrets

Patents

Infosec

Cybersecurity

Industrial Espionage



Follow

Written by Berend Watchus

26 followers · 45 following

Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former futurist event organizer & care innovation lab volunteer

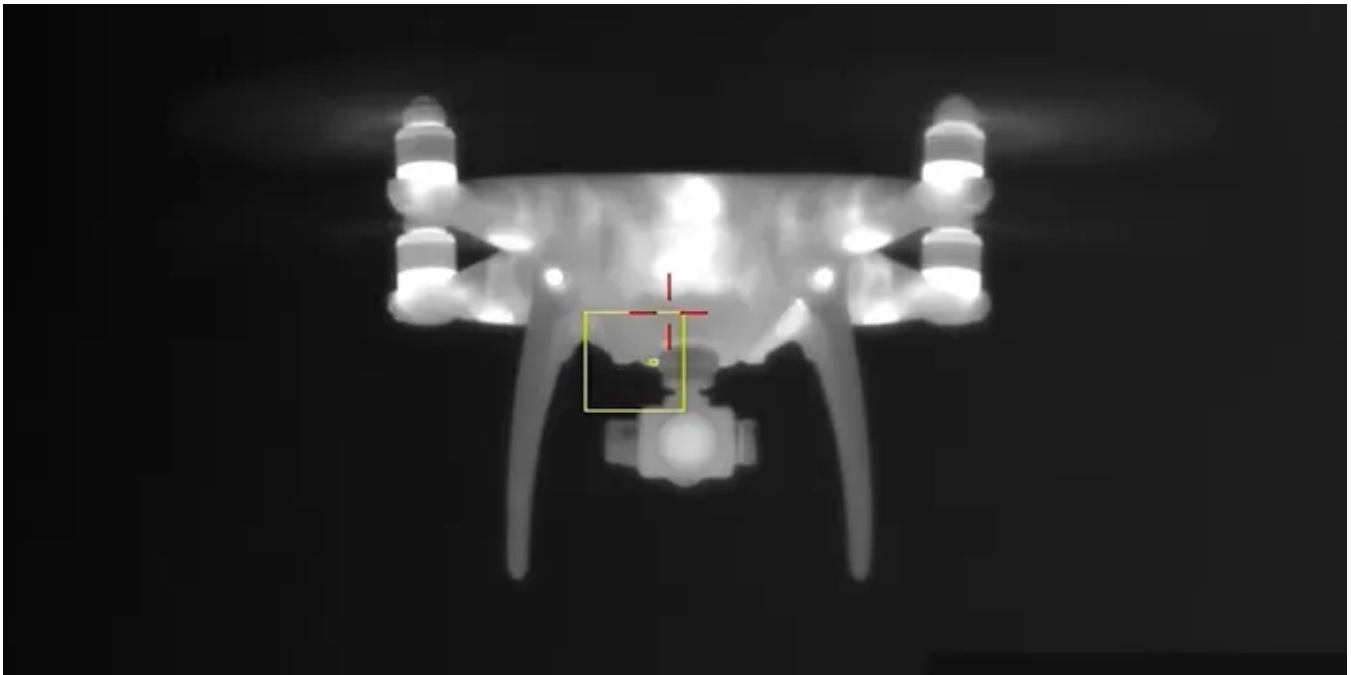
No responses yet



Write a response

What are your thoughts?

More from Berend Watchus



 In System Weakness by Berend Watchus

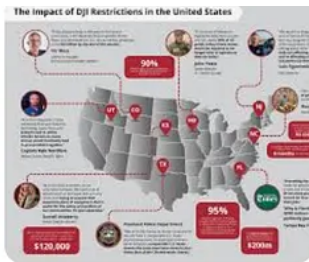
NATO Counter-Drone Operations in Belgium: Deep Technical Analysis for Cyber Security...

Author: Berend Watchus

Nov 12



 FDR1
De DJI-ban in de VS: protectionis...



 DroneDJ
This is what a DJI drone ban...



 Reddit
DJI drone sales ba...



 NEW CAMERA
USA DJI Drone Ban 2025: Save Y...



 Y.M.Cinema Magazine



 CineD



 YouTube



In System Weakness by Berend Watchus

The DJI Drone Ban: Comprehensive 2025 Update for Cybersecurity Professionals

Author: Berend Watchus includes free download in pdf

Nov 24



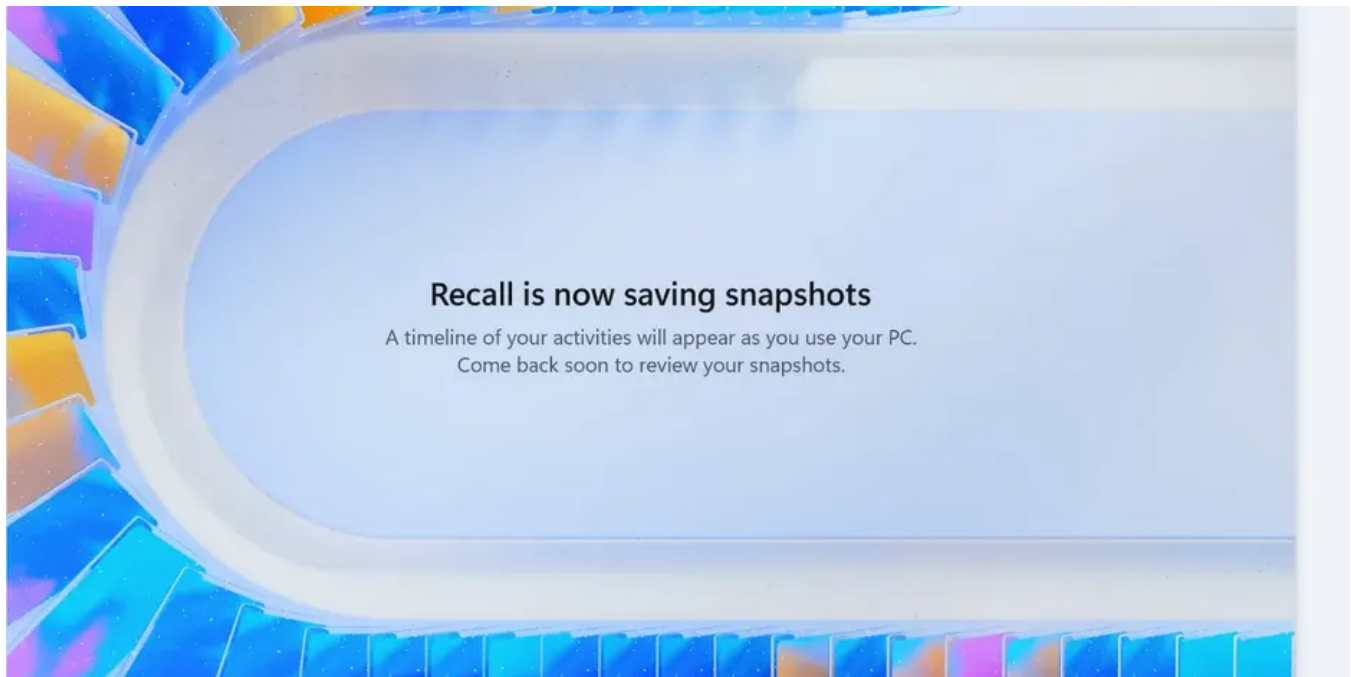
In System Weakness by Berend Watchus

Whisper Leak: Your LLM Chats Are Not as Private as You Think (Even Encrypted!)

Author: Berend Watchus November 11, 2025 Blog article

Nov 7





In System Weakness by Berend Watchus

Windows Recall: The VBS Enclave Paradox and the Evolution of Endpoint Espionage

Author: Berend Watchus November 11, 2025 Blog article

Nov 11  1



See all from Berend Watchus

Recommended from Medium



 Will Lockett 

Peter Thiel Just Revealed How Utterly Screwed The Entire AI Industry Is

The AI bubble is bursting.

🌟 Nov 22 🖱️ 9.6K 💬 266



:2510.01171v3 [cs.CL] 10 Oct 2025

ABSTRACT

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental, pervasive data-level driver: *typicality bias* in preference data, whereby annotators systematically favor familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce *Verbalized Sampling (VS)*, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., “Generate 5 jokes about coffee and their corresponding probabilities”). Comprehensive experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by $1.6\text{--}2.1\times$ over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity.

Problem: Typicality Bias Causes Mode Collapse

Tell me a joke about coffee.

Solution: Verbalized Sampling (VS) Mitigates Mode Collapse

Different prompts collapse to different modes:

1. Tell me a joke about coffee. 2. Tell me a story about coffee. 3. Tell me a poem about coffee. 4. Tell me a song about coffee. 5. Tell me a dance about coffee.

 In Artificial Intelligence in Plain English by Kiran Maan

Stanford Just Ended Prompt Engineering With a Single Sentence....

A new research method quietly restores the creativity we thought AI models had permanently lost... without retraining, plugins, or “advanced”...

🌟 ... 1K 29



ABSTRACT

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental, pervasive data-level driver: *typicality bias* in preference data, whereby annotators systematically favor familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce **Verbalized Sampling (VS)**, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., “Generate 5 jokes about coffee and their corresponding probabilities”). Comprehensive experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by $1.6\text{--}2.1\times$ over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity.

Problem: *Typicality Bias* Causes Mode Collapse

Tell me a joke about coffee

Solution: Verbalized Sampling (VS) Mitigates Mode Collapse

Different prompts collapse to different modes:

 In Generative AI by Adham Khaled

Stanford Just Killed Prompt Engineering With 8 Words (And I Can't Believe It Worked)

ChatGPT keeps giving you the same boring response? This new technique unlocks 2× more creativity from ANY AI model — no training required...

🌟 Oct 20 19.1K 468



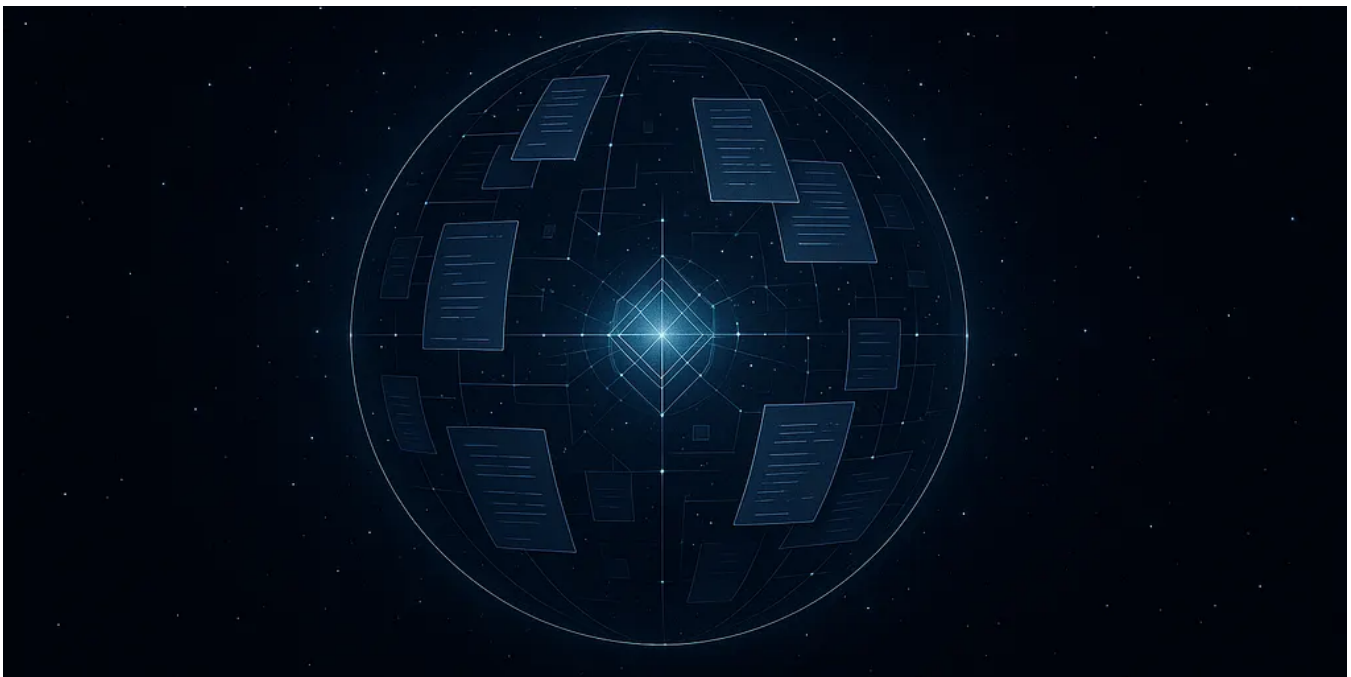


ichigo

I Accidentally Made Claude 45% Smarter. Here's How.

At 2 AM, I accidentally discovered psychological prompting tricks that sound insane but are backed by real research. Tested them on 40+...

Nov 21 🖱️ 1.2K 💬 26

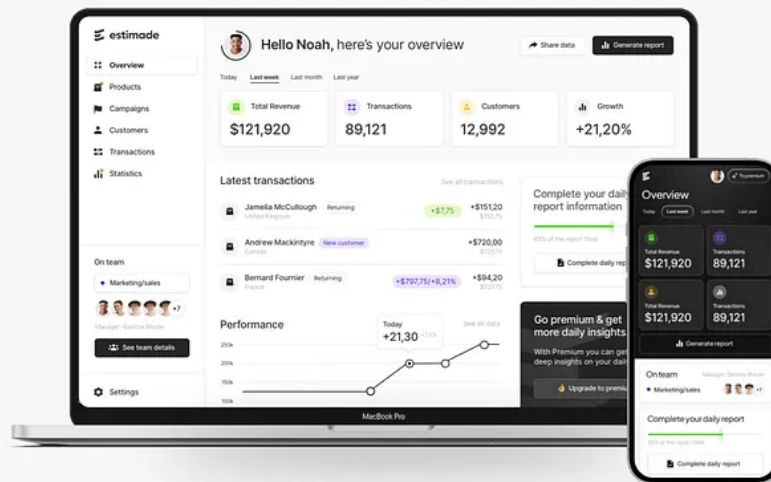


 stunspot

🔹🌐 Everyone Is Wrong About NotebookLM

Why Google's quiet little lab tool is the most important AI product nobody understands.

Nov 17 🖱️ 912 💬 32



 Michal Malewicz 📖

The End of Dashboards and Design Systems

Design is becoming quietly human again.

🌟 Nov 26 🖱️ 1.5K 💬 58



See more recommendations